On the Risk of Misinformation Pollution with Large Language Models

Year: 2023 | *Citations: 169* | *Authors: Yikang Pan, Liangming Pan, Wenhu Chen*

Abstract

In this paper, we comprehensively investigate the potential misuse of modern Large Language Models (LLMs) for generating credible-sounding misinformation and its subsequent impact on information-intensive applications, particularly Open-Domain Question Answering (ODQA) systems. We establish a threat model and simulate potential misuse scenarios, both unintentional and intentional, to assess the extent to which LLMs can be utilized to produce misinformation. Our study reveals that LLMs can act as effective misinformation generators, leading to a significant degradation in the performance of ODQA systems. To mitigate the harm caused by LLM-generated misinformation, we explore three defense strategies: prompting, misinformation detection, and majority voting. While initial results show promising trends for these defensive strategies, much more work needs to be done to address the challenge of misinformation pollution. Our work highlights the need for further research and interdisciplinary collaboration to address LLM-generated misinformation and to promote responsible use of LLMs.